# Introducción a ML y GenAI

**Árboles de Decisión (Decision Trees)**

Ariel Ramos Vela

01-10-2024

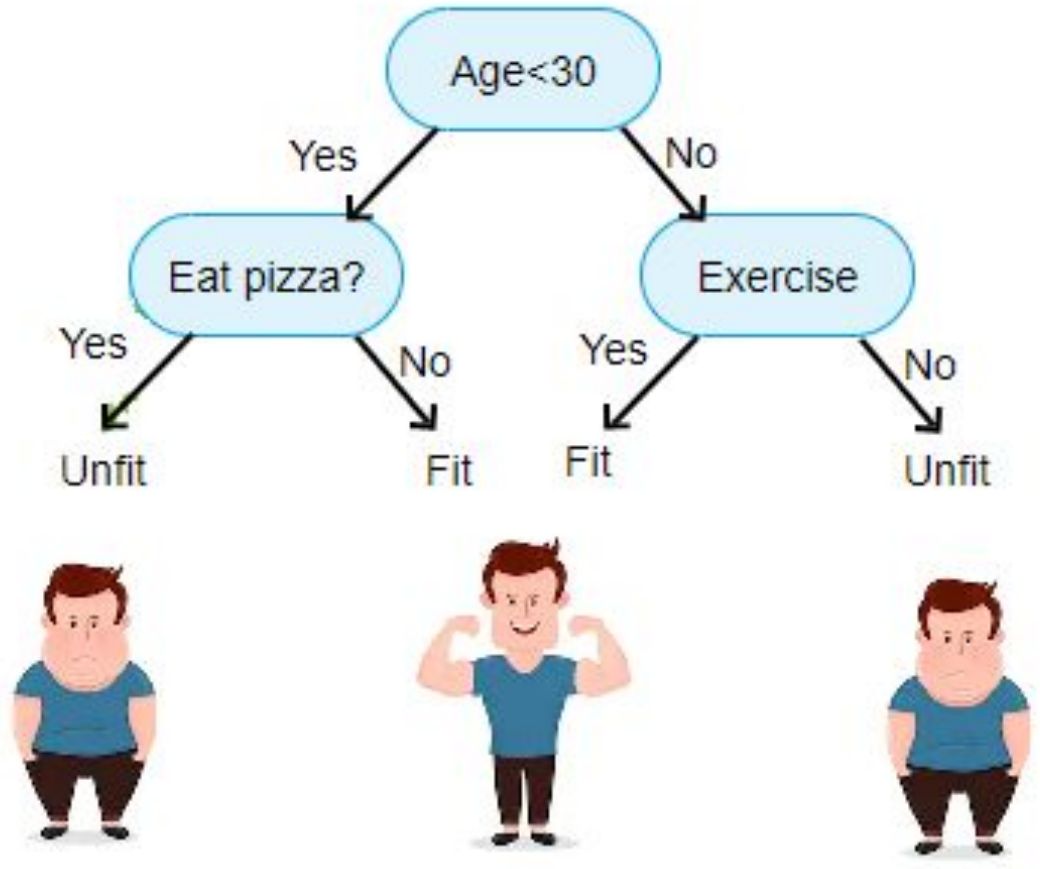
# Agenda

1. Introducción a los árboles de decision
2. Componentes de un árbol de decision
3. Algoritmos para construir árboles
4. Métricas y criterios de division
5. Ventajas y limitaciones
6. Taller 6

# ¿Qué es un árbol de decision (Decision Trees)?

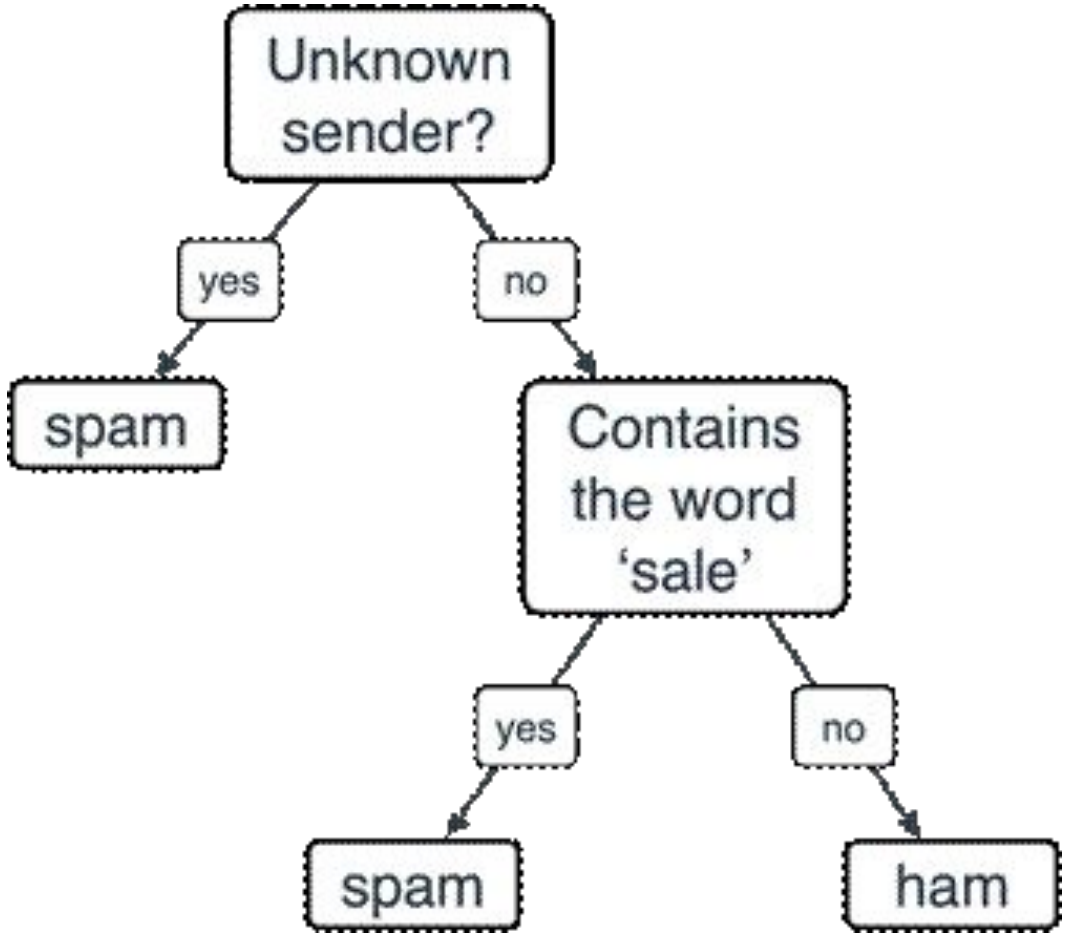
**Definición**: Un árbol de decisión es un modelo predictivo que divide iterativamente los datos en subconjuntos basados en características específicas.

Se utiliza tanto para **clasificación** como para regresión.



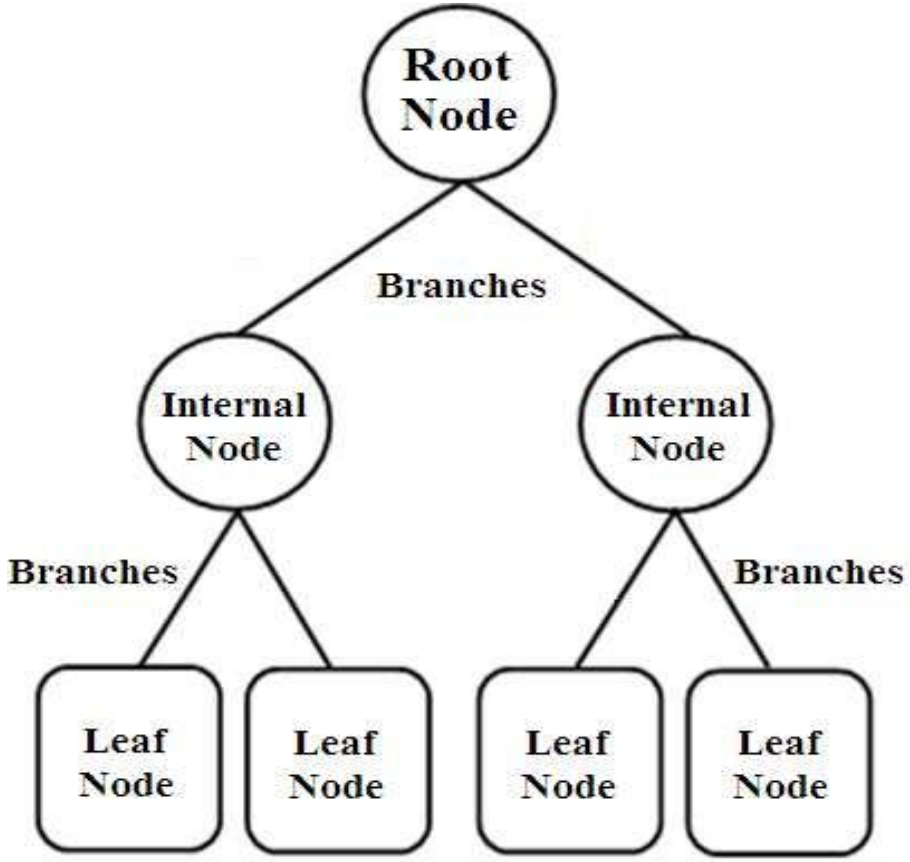
# Aplicaciones comunes

- Diagnóstico medico
- Análisis de crédito
- Filtrado de correo no deseado
- Predicciones sobre supervivencia (Titanic dataset)



# Componentes de un árbol de decisión

- **Nodos**: Preguntas o decisiones basadas en atributos.
- **Ramas (branches)**: Resultado de una decisión.
- **Hojas (leaf)**: Resultados finales o clases.
- **Raíz (root)**: Nodo inicial donde comienza la partición.

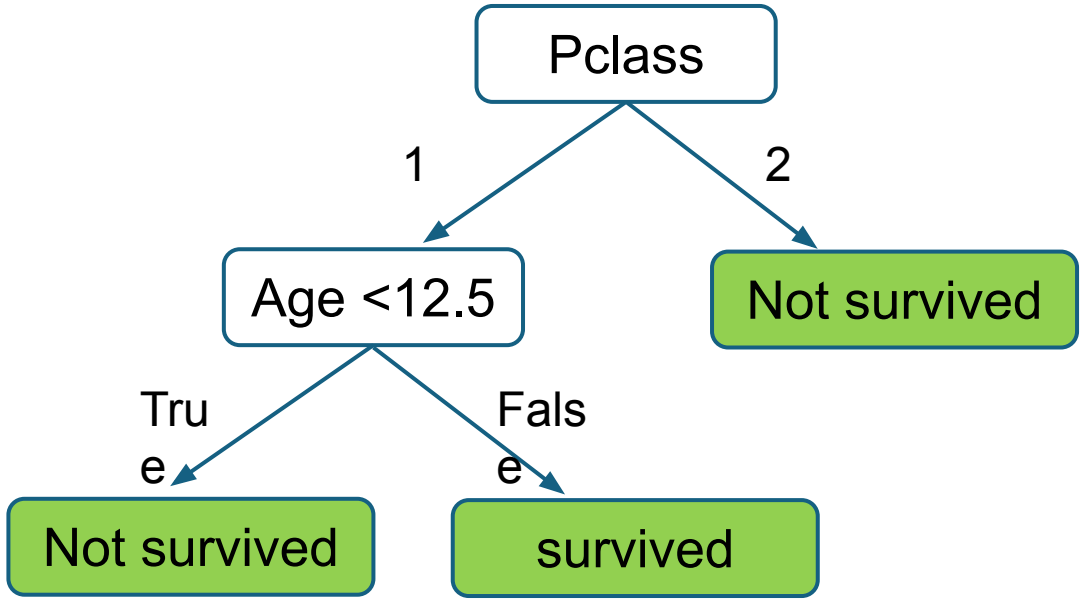


# Algoritmos para construir árboles

1. **ID3**: Usa Ganancia de Información (basada en entropía).
2. **CART**: Árboles de clasificación y regresión. Utiliza **Gini** o el Error Cuadrático Medio.
3. **C4.5**: Extensión de ID3 que admite variables continuas y podas (pruning).

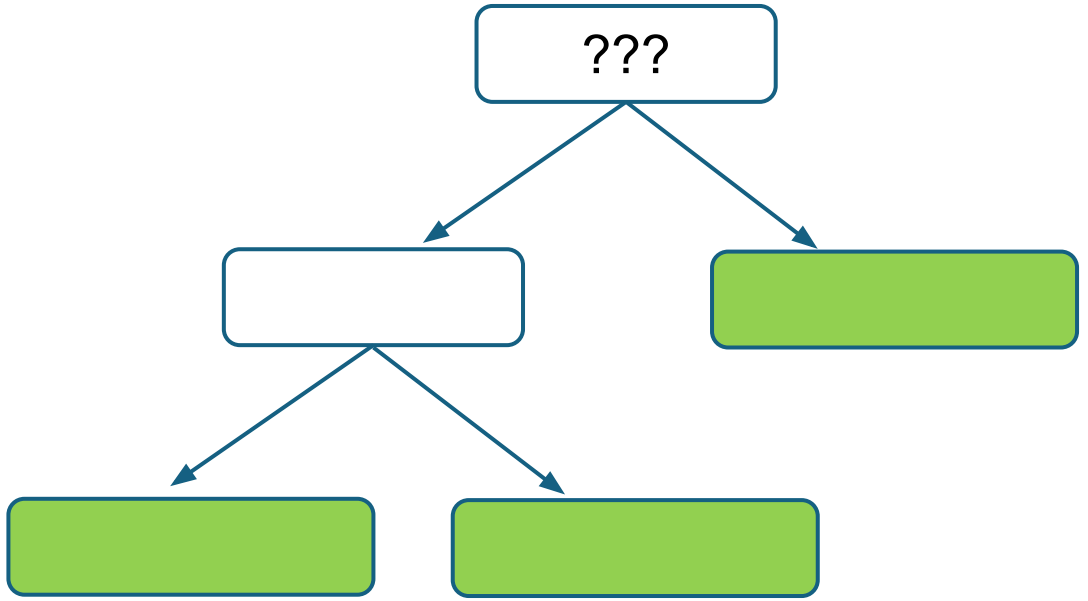
# Titanic dataset

| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | 7 | 0 |
| Male | 2 | 12 | 0 |
| Female | 1 | 18 | 1 |
| Female | 1 | 35 | 1 |
| Male | 1 | 38 | 1 |
| Male | 2 | 50 | 0 |
| Female | 2 | 83 | 0 |



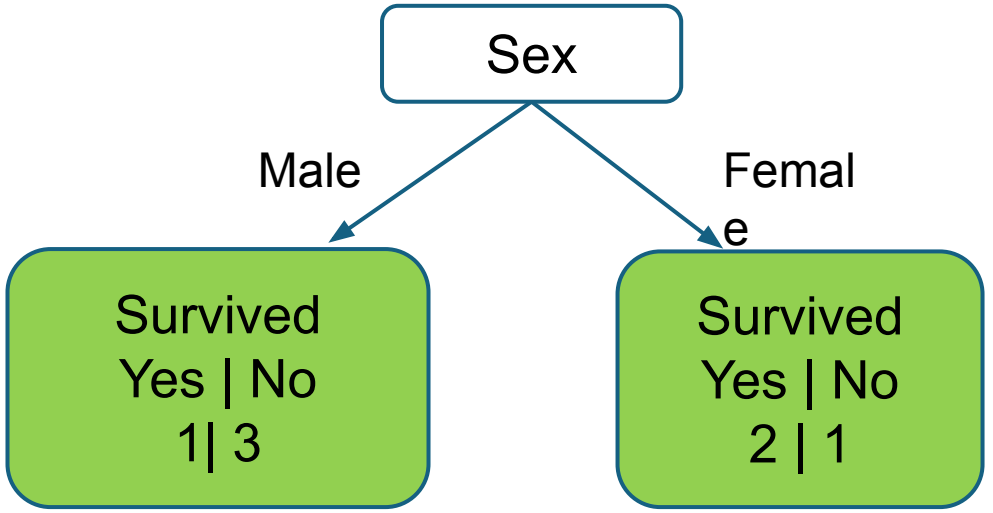
# Titanic dataset

| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | 7 | 0 |
| Male | 2 | 12 | 0 |
| Female | 1 | 18 | 1 |
| Female | 1 | 35 | 1 |
| Male | 1 | 38 | 1 |
| Male | 2 | 50 | 0 |
| Female | 2 | 83 | 0 |



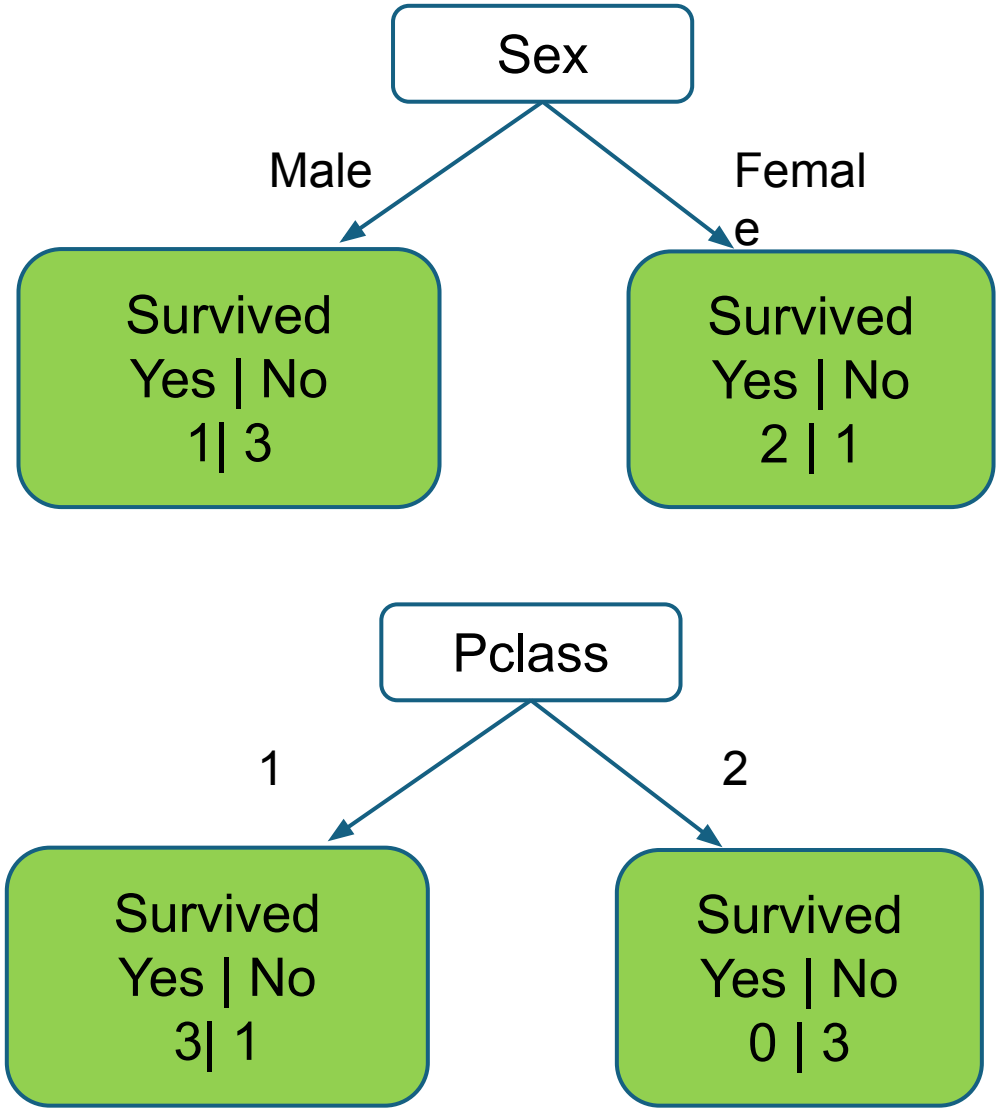
# Titanic dataset

| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | 7 | 0 |
| Male | 2 | 12 | 0 |
| Female | 1 | 18 | 1 |
| Female | 1 | 35 | 1 |
| Male | 1 | 38 | 1 |
| Male | 2 | 50 | 0 |
| Female | 2 | 83 | 0 |



# Titanic dataset

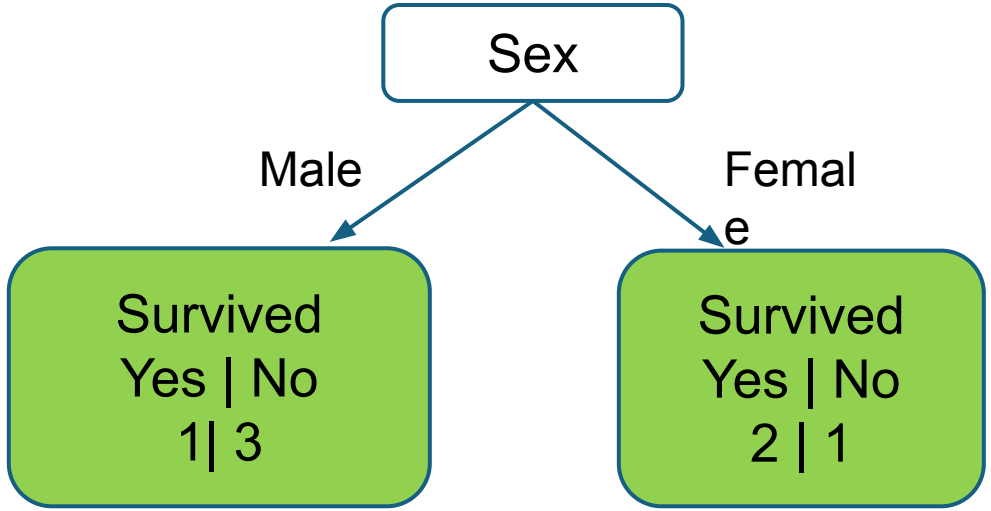
| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | 7 | 0 |
| Male | 2 | 12 | 0 |
| Female | 1 | 18 | 1 |
| Female | 1 | 35 | 1 |
| Male | 1 | 38 | 1 |
| Male | 2 | 50 | 0 |
| Female | 2 | 83 | 0 |



# Titanic dataset

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

$$Impuridad\ Gini\ de\ un\ nodo = 1\ - p(yes)^2 - p(no)^2$$




$$= 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2$$

$$= 1 - (0.25)^2 - (0.75^2)$$

$$= 0.375$$

$$= 1 - \left(\frac{2}{2+1}\right)^2 - \left(\frac{1}{2+1}\right)^2$$

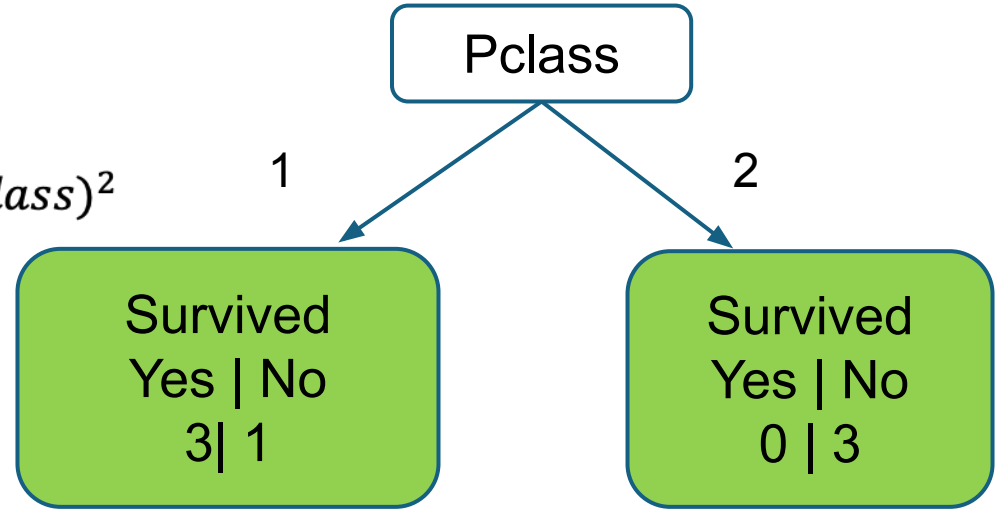
$$= 0.444$$

$$Impuridad\ Gini\ total: promedio\ ponderado\ (weighted\ average)\ de\ las\ ramas$$

$$= \left(\frac{4}{4+3}\right)(0.375) + \left(\frac{3}{4+3}\right)(0.444) = 0.405$$

# Titanic dataset

$$Impuridad\ Gini\ de\ un\ nodo = 1\ - p(class\_1)^2 - p(class)^2$$




$$= 1 - \left(\frac{3}{1+3}\right)^2 - \left(\frac{1}{1+3}\right)^2$$

$$= 1 - (0.75)^2 - (0.25^2)$$

$$= 0.375$$

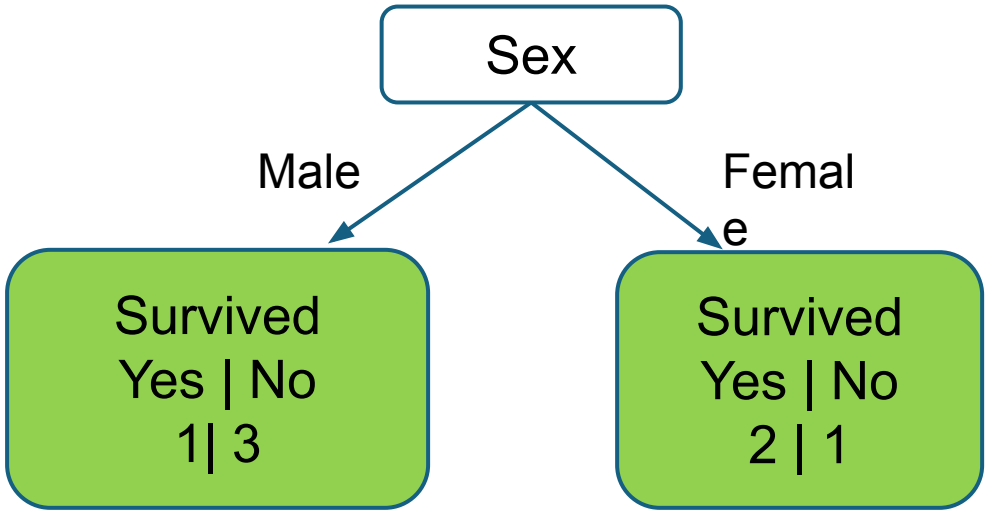
$$= 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2$$

$$= 0$$

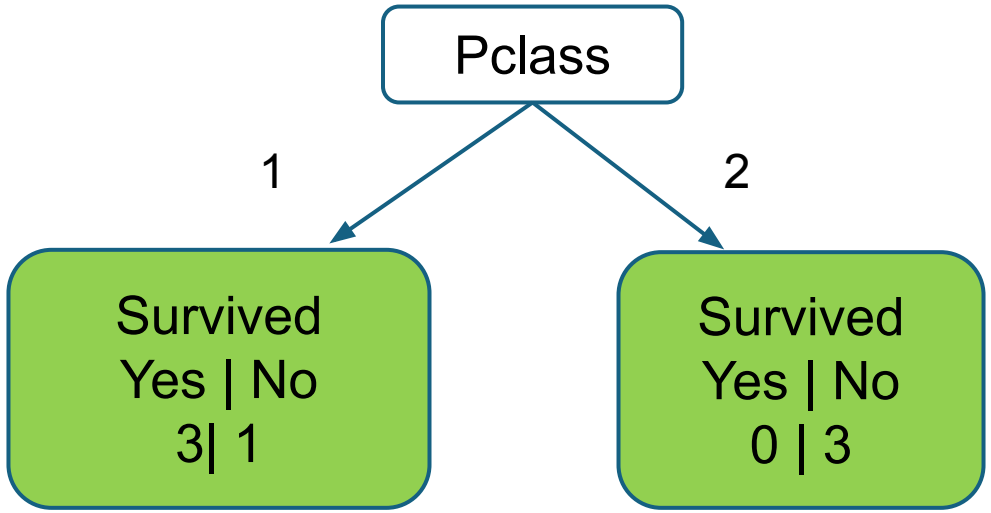
$$Impuridad\ Gini\ total: promedio\ ponderado\ (weighted\ average)\ de\ las\ ramas$$

$$= \left(\frac{4}{4+3}\right)(0.375) + \left(\frac{3}{4+3}\right)(0) = 0.214$$

# Titanic dataset



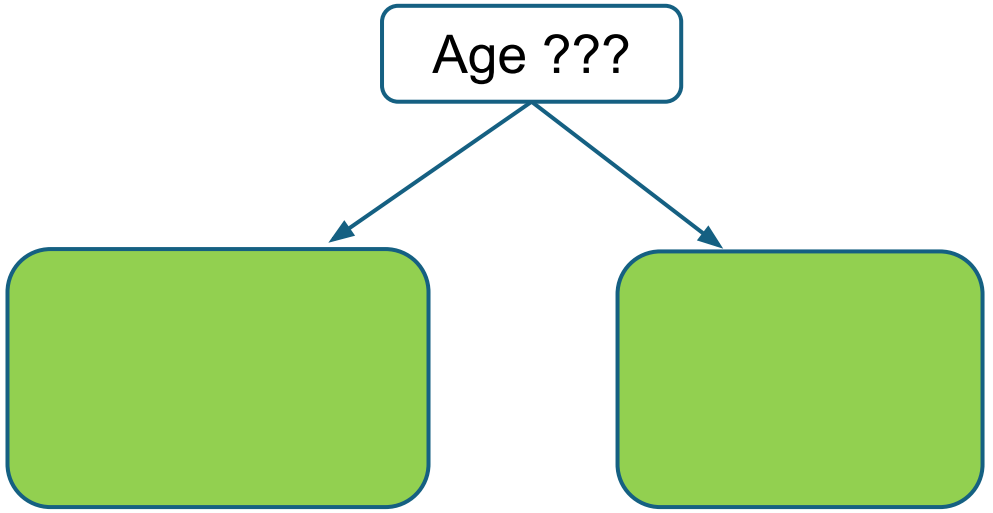

0.405




0.214

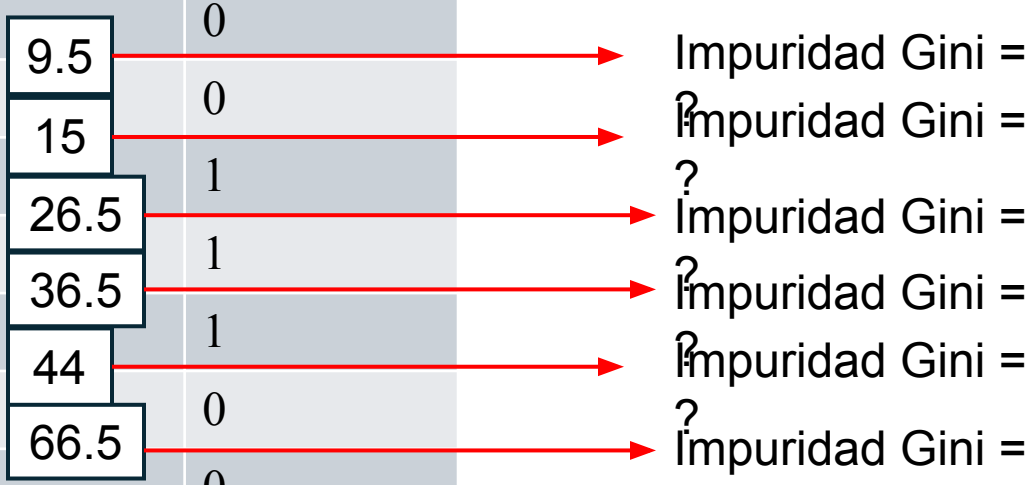
# Titanic dataset

| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | **7** | 0 |
| Male | 2 | **12** | 0 |
| Female | 1 | **18** | 1 |
| Female | 1 | **35** | 1 |
| Male | 1 | **38** | 1 |
| Male | 2 | **50** | 0 |
| Female | 2 | **83** | 0 |



# Titanic dataset

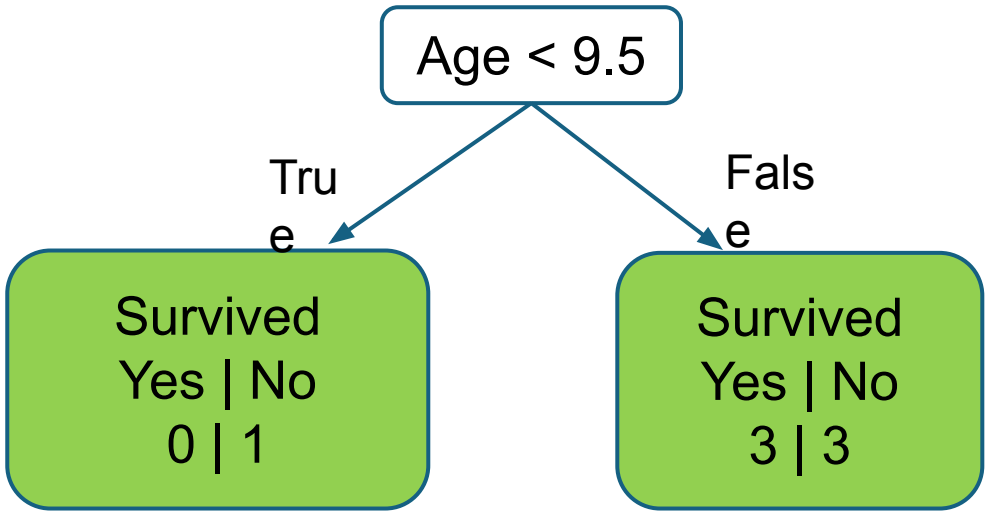
| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | **7** | 0 |
| Male | 2 | **12** | 0 |
| Female | 1 | **18** | 1 |
| Female | 1 | **35** | 1 |
| Male | 1 | **38** | 1 |
| Male | 2 | **50** | 0 |
| Female | 2 | **83** | 0 |



# Titanic dataset

| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | **7** | 0 |
| Male | 2 | **12** | 0 |
| Female | 1 | **18** | 1 |
| Female | 1 | **35** | 1 |
| Male | 1 | **38** | 1 |
| Male | 2 | **50** | 0 |
| Female | 2 | **83** | 0 |

9.5
15
26.5
36.5
44
66.5




Impuridad Gini = $1 - p(yes)^2 - p(no)^2$

$= 1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2$

$= 0$

$= 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2$

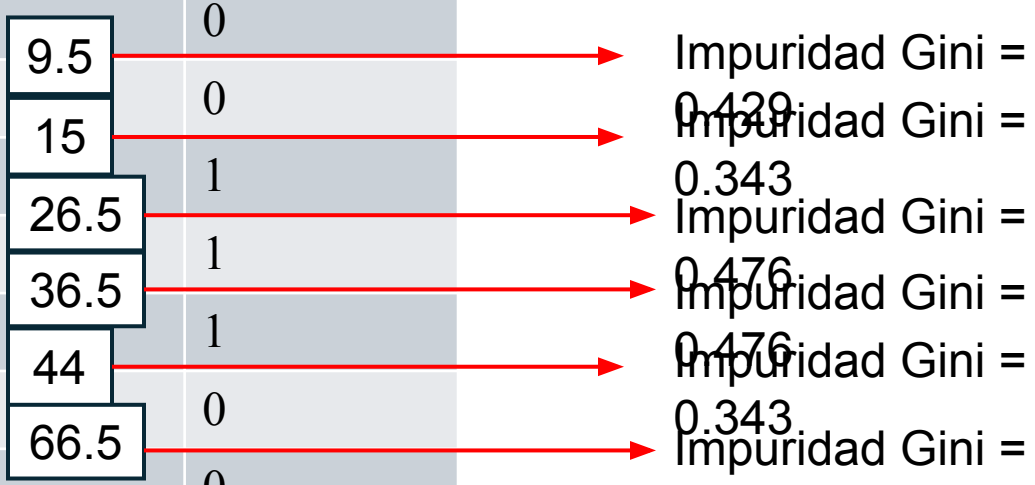
$= 1 - 0.25 - 0.25 = 0.5$

$$weightd\ average = \left(\frac{1}{7}\right) * 0 + \left(\frac{6}{7}\right) * 0.5$$

$= 0.429$

# Titanic dataset

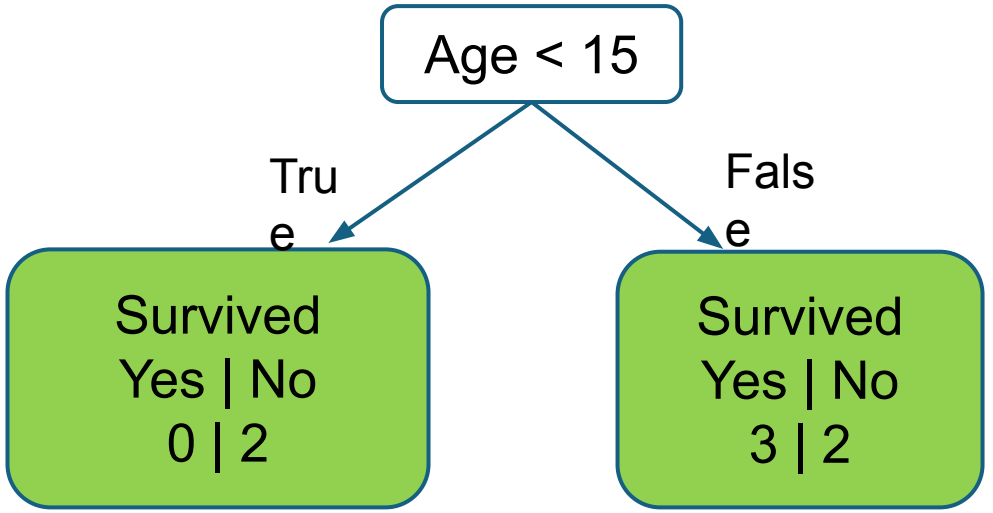
| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | **7** | 0 |
| Male | 2 | **12** | 0 |
| Female | 1 | **18** | 1 |
| Female | 1 | **35** | 1 |
| Male | 1 | **38** | 1 |
| Male | 2 | **50** | 0 |
| Female | 2 | **83** | 0 |



# Titanic dataset

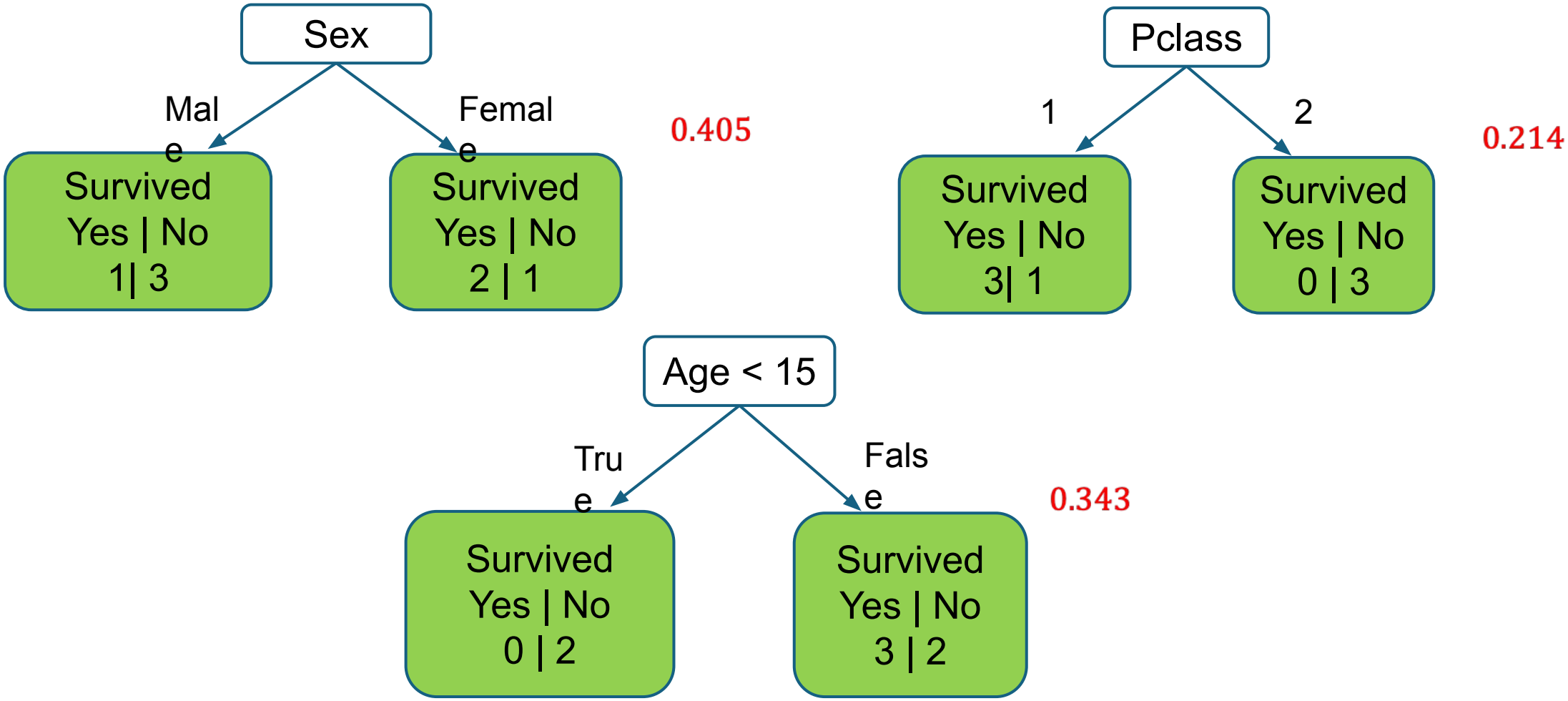
| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | **7** | 0 |
| Male | 2 | **12** | 0 |
| Female | 1 | **18** | 1 |
| Female | 1 | **35** | 1 |
| Male | 1 | **38** | 1 |
| Male | 2 | **50** | 0 |
| Female | 2 | **83** | 0 |

9.5
15
26.5
36.5
44
66.5



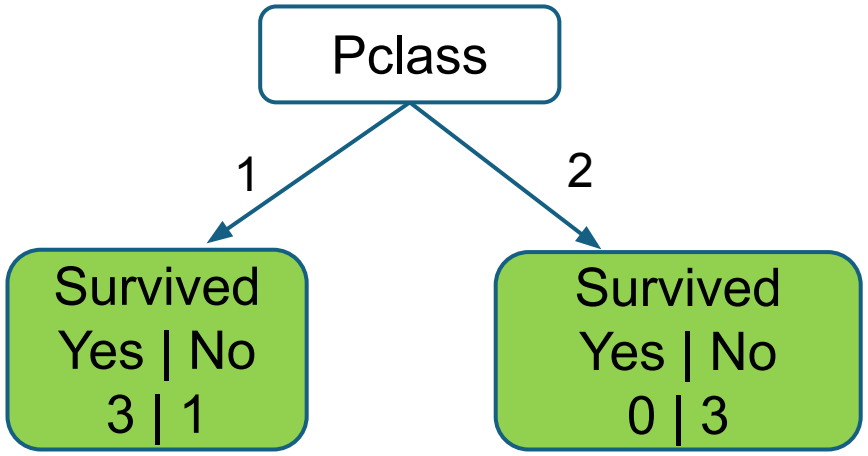

0.343

# Titanic dataset



# Titanic dataset

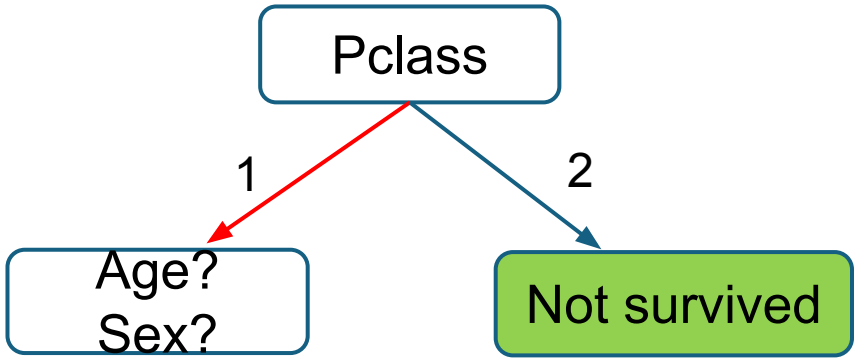
| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | 7 | 0 |
| Male | 2 | 12 | 0 |
| Female | 1 | 18 | 1 |
| Female | 1 | 35 | 1 |
| Male | 1 | 38 | 1 |
| Male | 2 | 50 | 0 |
| Female | 2 | 83 | 0 |




This node is impure

# Titanic dataset

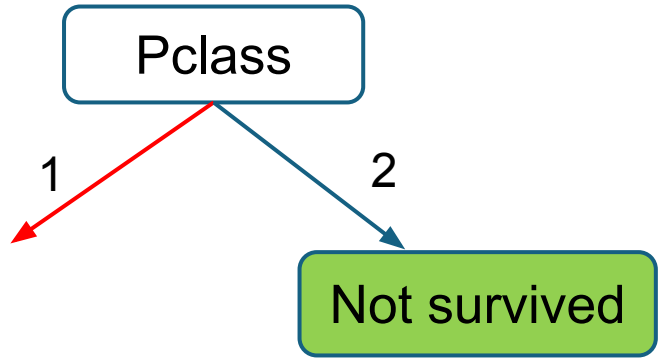
| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | 7 | 0 |
| Male | 2 | 12 | 0 |
| Female | 1 | 18 | 1 |
| Female | 1 | 35 | 1 |
| Male | 1 | 38 | 1 |
| Male | 2 | 50 | 0 |
| Female | 2 | 83 | 0 |

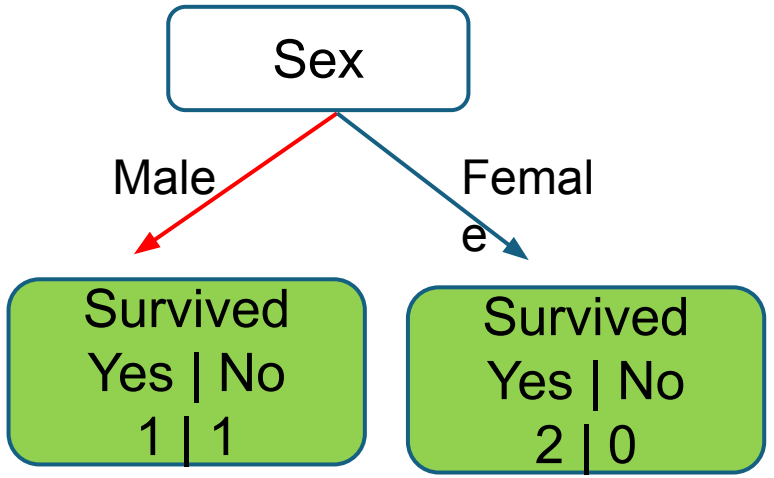



Mismo proceso, pero con un subset de datos.

# Titanic dataset

| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | 7 | 0 |
| Male | 2 | 12 | 0 |
| Female | 1 | 18 | 1 |
| Female | 1 | 35 | 1 |
| Male | 1 | 38 | 1 |
| Male | 2 | 50 | 0 |
| Female | 2 | 83 | 0 |




Impureza Gini:
0.25

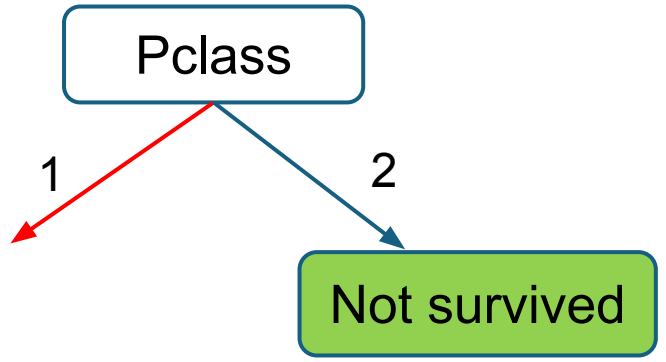
# Titanic dataset

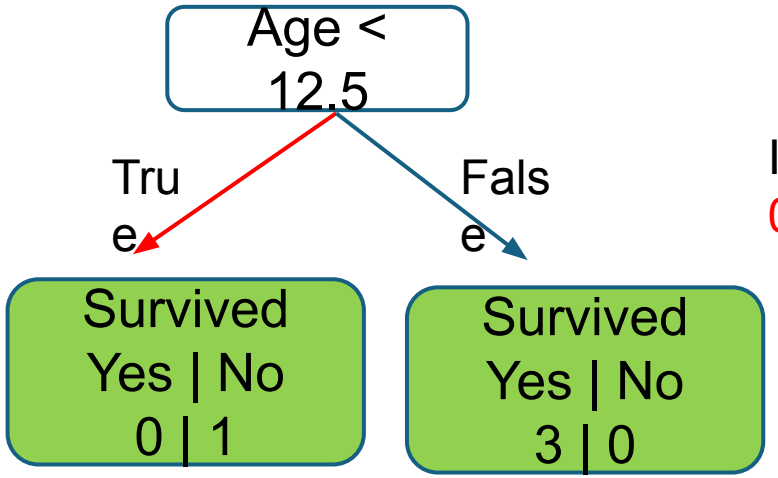
| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | 7 | 0 |
| Male | 2 | 12 | 0 |
| Female | 1 | 18 | 1 |
| Female | 1 | 35 | 1 |
| Male | 1 | 38 | 1 |
| Male | 2 | 50 | 0 |
| Female | 2 | 83 | 0 |

12.5

26.5

36.5

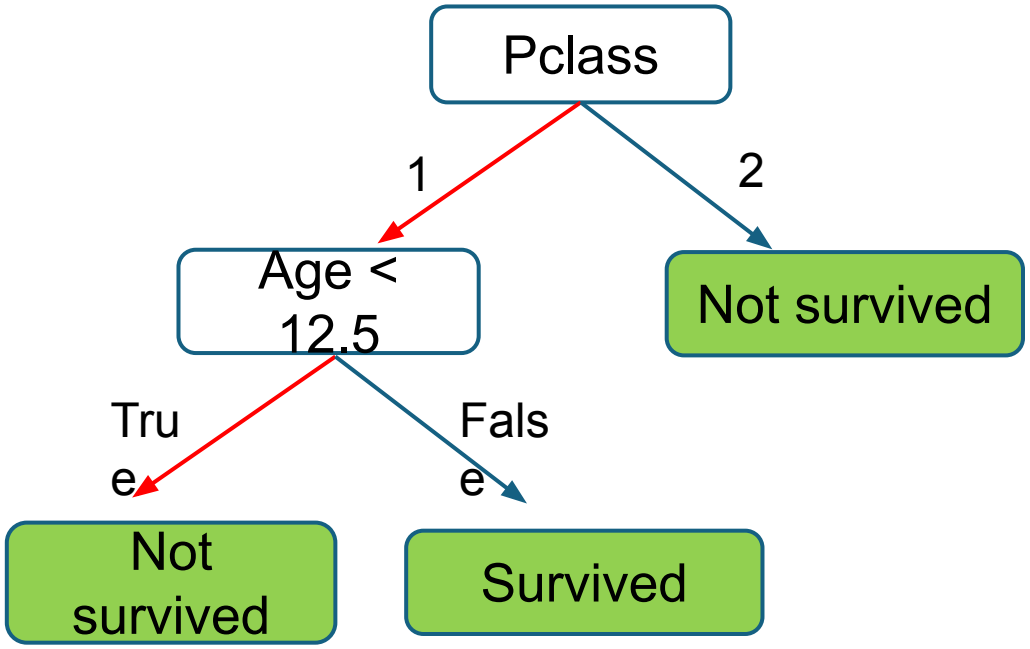



Impureza Gini:
0

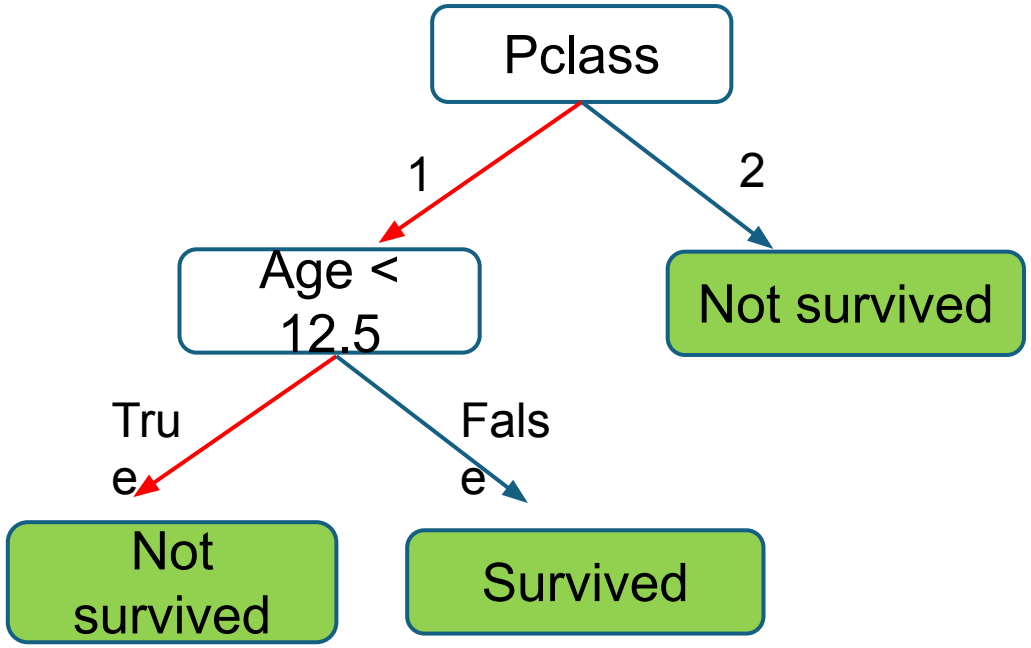
# Titanic dataset

| Sex | Pclass | Age | Survived |
|---|---|---|---|
| Male | 1 | 7 | 0 |
| Male | 2 | 12 | 0 |
| Female | 1 | 18 | 1 |
| Female | 1 | 35 | 1 |
| Male | 1 | 38 | 1 |
| Male | 2 | 50 | 0 |
| Female | 2 | 83 | 0 |



# Predecir nuevos valores

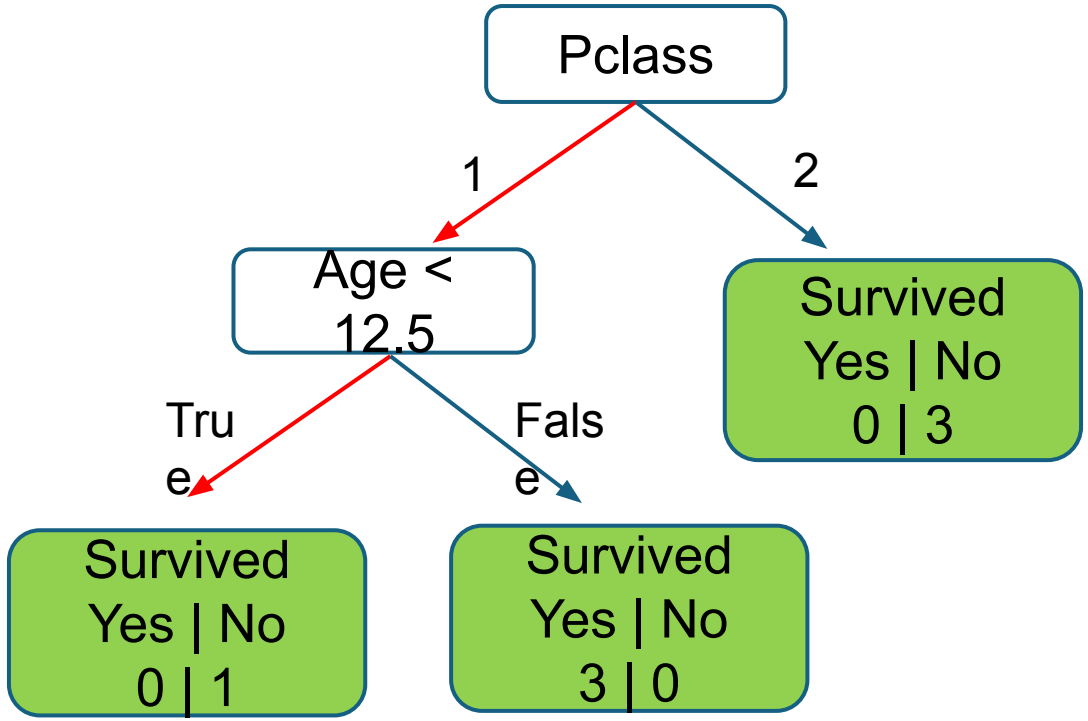
| Sex | Pclass | Age |
|---|---|---|
| Female | 1 | 26 |




Predicción:
**Survived**

# Consideraciones

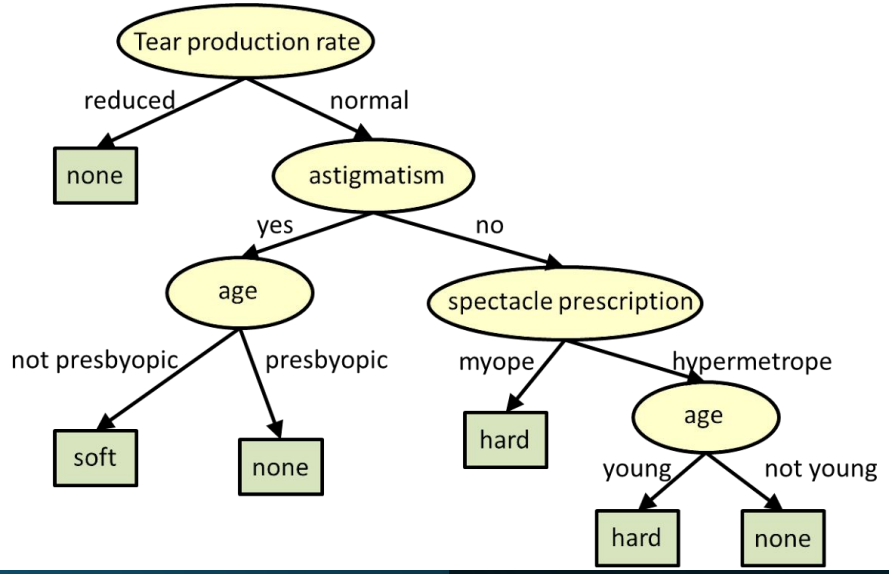
- Si tenemos pocos datos en un node hoja (ejemplo 1), es probable que tengamos sobreajuste (**overfitting)**.
- Soluciones:
  - Poda de árboles (Pruning)
  - Determinar un número mínimo de nodos (hyperparametro) para ser considerado **leaf.**

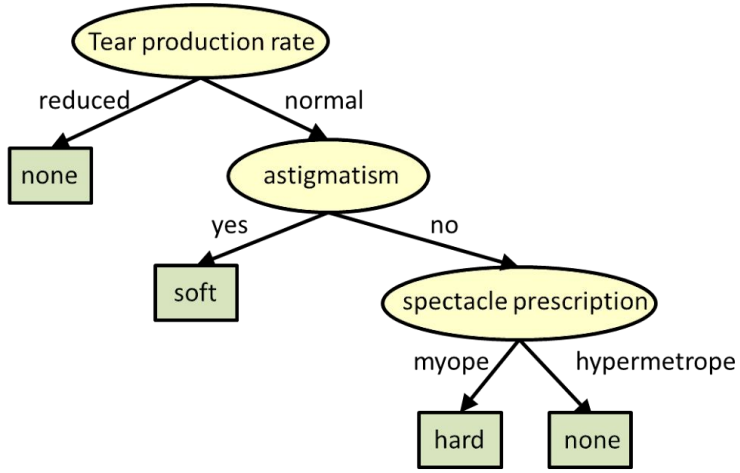


# Pruning

- **Definición**:
  La poda es una técnica utilizada para **reducir el tamaño del árbol** de decisión eliminando ramas que proporcionan poca o ninguna mejora en la predicción.

- **Tipos:**
  - Pre-pruning (poda anticipada)
  - Post-pruning (poda posterior)



# Ventajas de los árboles de decisión

- Interpretación fácil y visual.
- No requieren normalización de los datos.
- Funcionan bien con datos categóricos y numéricos.
- Pueden manejar datos con valores faltantes.

# Limitaciones de los árboles de decisión

- **Sobreajuste**: Los árboles muy profundos tienden a memorizar los datos.
- **Inestabilidad**: Pequeños cambios en los datos pueden crear un árbol completamente diferente.
- **Podas (Pruning)**: Necesarias para mejorar la generalización.

# Mejora de los árboles de decisión

- **Random Forest**: Combinación de múltiples árboles para reducir la variabilidad.
- **Gradient Boosting**: Ensamblaje de árboles secuenciales para mejorar la precisión.